Tracking System with Re-identification Using a RGB String Kernel

Amal Mahboubi, Luc Brun, Donatello Conte, Pasquale Foggia, Mario Vento

Tracking System with Re-identification using a RGB String Kernel

Amal Mahboubi[1], Luc Brun[1],
Donatello Conte[2], Pasquale Foggia[3], and Mario Vento[3]

[1] GREYC UMR CNRS 6072, Equipe Image ENSICAEN
6, boulevard Maréchal Juin F-14050 Caen, FRANCE,
amal.mahboubi@unicaen.fr , luc.brun@ensicaen.fr

[2] Université François-Rabelais de Tours, LI EA 6300
64, Avenue Jean Portalis, F-37200, Tours, France
donatello.conte@univ-tours.fr

[3] Dipartimento di Ingegneria dell'Informazione,
Ingegneria Elettrica e Matematica Applicata
Università di Salerno, Via Ponte Don Melillo, 1 I-84084 Fisciano (SA), ITALY
pfoggia@unisa.it , mvento@unisa.it

Abstract. People re-identification consists to identify a person which comes back in a scene where it has been previously detected. This key problem in visual surveillance applications may concern single or multi camera systems. Features encoding each person should be rich enough to provide an efficient re-identification while being sufficiently robust to remain significant through the different phenomena which may alter the appearance of a person in a video. We propose in this paper a method which encodes people's appearance through a string of salient points. The similarity between two such strings is encoded by a kernel. This last kernel is combined with a tracking algorithm in order to associate a set of strings to each person and to measure similarities between persons entering into the scene and persons who left it.

Keywords: Re-identification, String kernel, Visual surveillance

1 Introduction

The purpose of re-identification is to identify people coming back onto the field of view of a camera. Several types of features including interest point [9, 2], histograms [3, 10, 16, 8], shape [6], graph based representation [17, 11, 13, 2], have been proposed in the literature. However, some features like histograms do not encode any spatial information while some others like interest point and graph based representations may induce a matching step which requires important execution times. Moreover complex features like bags or graphs [13] of interest

points, RAG [2] may be sensitive to the evolution of the appearance of a person in a video due to his displacements or occlusions.

Independently of the type of features used to perform the re identification step, re-identification methods may be split into two categories: methods of the first group [9] compute a unique signature for each object and perform the re-identification based on this single signature. Methods of the second group [3, 18] delay the re-identification which is then performed on a set of signatures. This last approach imposes to base the comparison between two objects on a comparison of two sets of signatures rather than between two individual signatures. However such an approach can potentially better capture the variability of the appearance of a person over a video.

Our approach belongs to the second category and describes the appearance of a person by a set of RGB string descriptors (Section 2) computed over a sliding window. A kernel between two sets of strings (Section 3) is then applied in order to encode the similarity between two persons. The integration of this kernel into a tracking method is described in Section 4 while Section 5 reports several experiments which demonstrate the validity of our approach.

2 RGB String Descriptor Construction Scheme

One of the main challenges for people re-identification is to capture peoples' appearance properties. As mentioned in Section 1, several modelings have been developed. However, although complex models such as graph based representation offer the advantage of a precise modeling of the object, they usually require a complex matching step and important execution times. An alternative solution consists in using a string descriptor. Indeed, string allows an effective comparison while preserving useful information of the region of interest. Although a string usually encodes less information than a graph we expect a greater stability of this simpler structure over time.

The first step of our method consists in separating subjects from the background. To that end, we use binary object masks [2] defined by a foreground detection with shadow removals. Each moving person within a frame is thus associated to a mask that we characterize using a salient string. Each character of this string is defined by the couple of its coordinates (x,y) and its RGB value. The construction of a salient string is outlined in Figure 1. This construction consists of the following 3 steps: First, we apply a Deriche edge detector on each moving person according to its binary mask within a frame.

The second stage consists in building a discriminating curve of the object. Contour points provided by the Deriche detector are used to build this curve. Let us consider the bounding box of an object obj_a. Thanks to the Deriche filter obj_a should be delineated by two main contours (Figure 1-step-1). For each height value y, the mean on the bounding box of the x coordinates weighted by the squared mean of the gradient should thus provide a central point inside obj_a (a character of the string). This average coordinate is described by the equation

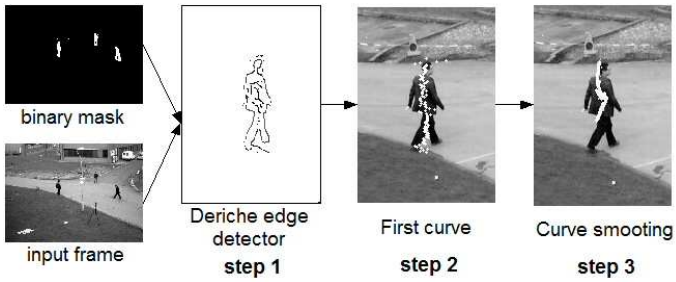

Fig. 1. RGB string construction steps

below:

$$\bar{x}(y) = \frac{\sum_{i=1}^{n} |\nabla I(x_i, y)|^2 . x_i}{\sum_{i=1}^{n} |\nabla I(x_i, y)|^2} \tag{1}$$

where (x_i, y) are the coordinates of the considered point and $|\nabla I(x_i, y)|$ is the amplitude of its gradient. The symbol n denotes the height of the bounding box.

The last and third step, enforce the quality of the resulting curve. Indeed, the resulting curve (Figure 1-step-2) is sensible to small perturbations of the gradient on each mask and contains important discontinuities. This last point may alter the similarity of two curves of a same person taken on two different frames. We thus propose to regularize this curve through an energy minimization framework. Our energy function (equation 2) combines two terms: the first one encodes the attachment to the initial curve $(\bar{x}_y)$. The second term is a regularization term which enforces the continuity of the curve.

We hence assume that, the energy functional is described as follows:

$$\jmath(x) = \sum_{y=1}^{n} (\bar{x}_y - x_y)^2 + \lambda (x_y - x_{y-1})^2 \tag{2}$$

where λ is a tuning parameter. The average coordinate $\bar{x}_y$ is given by equation 1 and x_y is the corresponding final coordinate. The symbol n denotes the height of the bounding box.

Minimization of equation 2 leads to search for the zeros of its gradient:

$$\frac{\partial \jmath}{\partial x_y} = 2(1 + 2\lambda)x_y - 2\lambda(x_{y-1} + x_{y+1}) - 2\bar{x}_y = 0 \tag{3}$$

Equation 3 corresponds to the formulation of a tridiagonal system which can be solved in $\mathcal{O}(n)$.

This last minimization step obtained using equation 3 provides the final curve, where each point is associated to its RGB value (Figure 1-step-3).

3 People description

Curves encoding people's appearance may be altered by the addition of erroneous extremities encoding for example a part of the floor or a difference of sampling due to the variations of the distance between a person and the camera. In order to cope with such variations we consider each curve as a string and encode the similarity between two strings thanks to the global alignment kernel defined by [4]:

$$K_{GA}(s_1, s_2) = \sum_{\pi \in A(n,m)} e^{-D_{s_1,s_2}(\pi)} \tag{4}$$

where n and m, denote the length of the first string s_1 and the second string s_2 respectively. An alignment is noted π and $A(n, m)$ represents the set of all alignments between s_1 and s_2. The symbol D denotes the Dynamic Time Warping distance. It measures the discrepancy between two strings s_1 and s_2 according to an alignment π. Function D is defined [4] as:

$$D_{s_1,s_2}(\pi) = \sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}) \tag{5}$$

where $s_1 = (x_i)_{i\in\{1,..,n\}}$, $s_2 = (y_i)_{i\in\{1,...m\}}$ and function φ corresponds to a distance function defined [4] as follows:

$$\varphi(x, y) = \frac{1}{2\sigma^2}\|x - y\|^2 + \log(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}) \tag{6}$$

where x and y denote the RGB values of the first object and the second object respectively. Symbol σ denotes a tuning parameter. The log term is added to the squared Euclidean distance $\|x - y\|^2$ in order to ensure the definite positiveness of K_{GA} (equation 4) [4].

3.1 People's Kernel

As the appearance of a person evolves in a scene, due to slight changes of the pose, the use of a single string is inappropriate to identify a person. Assuming that, the appearance of a person is established on a set of successive frames. The global appearance of a person over a video is described by a set of salient strings. The temporal window over which this set is built is called the history tracking window (HTW).

Each person in the video is hence not defined by a single string but by a set of strings (on HTW). This set may include outlier strings, due to slight changes of the pose or occlusion. The construction of a representative string based on a simple average of all the strings of a set (in the Hilbert space defined by the kernel) may be sensible to such outliers. We thus suggest to enforce the robustness of our representative string through the use of an one class SVM classifier in order to remove outliers.

Let $\mathcal{H}$ denotes the Hilbert space defined by K_{GA} (equation 4). In order to get a robust model encoding the mean appearance of a person, we first use K_{GA} to project the mapping of all strings onto the unit-sphere of $\mathcal{H}$. This operation is performed by normalizing our kernel [5]. Following [5], we then apply a one class ν-SVM on each set of strings describing a person. From a geometrical point of view, this operation is equivalent to model the set of projected strings by a spherical cap defined by a weight vector w and an offset ρ both provided by the ν-SVM algorithm. These two parameters define the hyper plane whose intersection with the unit sphere defines the spherical cap. Strings whose projection on the unit sphere lies outside the spherical cap are considered as outliers. Each person is thus encoded by a triplet (w, ρ, S) where S corresponds to the set of strings and (w, ρ) are defined from a one class ν-SVM. Figure 2 gives the geometric interpretation of (w, ρ); The parameter w indicates the center of the spherical cap and may be intuitively understood as the vector encoding the mean appearance of a person over its HTW window. The parameter ρ influence the radius of the spherical cap and may be understood as the extend of the set of representative strings in S.

Let $P_A = (w_A, \rho_A, S_A)$ and $P_B = (w_B, \rho_B, S_B)$ denote two triplets encoding two persons A and B. The distance between A and B is defined from the angle between vectors w_A and w_B defined as follows [5]:

$$d_{sphere}(w_A, w_B) = arccos\left(\frac{w_A^T K_{A,B} w_B}{\|w_A\|\|w_B\|}\right) \tag{7}$$

where $\|w_A\|$ and $\|w_B\|$ denote the norms of w_A and w_B in $\mathcal{H}$ and $K_{A,B}$ is a $|S_A| \times |S_B|$ matrix defined by $K_{A,B} = (K_{norm}(t, t'))_{(t,t') \in S_A \times S_B}$, where K_{norm} denotes our normalized kernel.

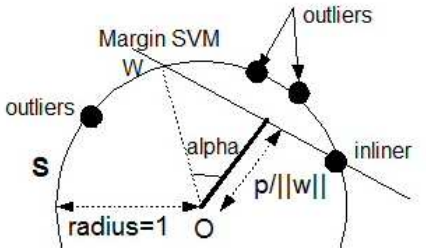

Fig. 2. Geometrical interpretation of equation 7

Based on d_{sphere}, the kernel between A and B is defined as the following product of RBF kernels:

$$K_{change}(P_A, P_B) = e^{\frac{-d^2_{sphere}(w_A, w_B)}{2\sigma^2_{moy}}} e^{\frac{-(\rho_A - \rho_B)^2}{2\sigma^2_{origin}}} \tag{8}$$

Where σ_{moy} and σ_{origin} are tuning variables.

4 Re-identification

Our tracking algorithm is based on a previous work [13]. The tracking algorithm uses four labels 'new', 'get-out', 'unknown' and 'get-back' with the following meaning: '*new*' refers to an object classified as new, '*get-out*' represents an object leaving the scene, '*unknown*' describes a query object (an object recently appeared, not yet classified) and '*get-back*' refers to an object classified as an old one after a re-identification step.

Re-identification is achieved using the similarity (equation 8) between each unknown person and all the get-out persons. All masks detected in the first frame of a video are regarded as new persons. Then a mask detected in frame $t+1$ is regarded as matched if a large overlap exists between its bounding box and a bounding box defined in frame t. In this case, the mask is assigned to the same person than in frame t.

If one mask at frame t does not have any successor in frame $t+1$, the associated person is marked as get-out. Its curve at frame t is added to the sliding HTW window containing the last strings of this person. Its triplet $P = (w, \rho, S)$ (Section 3) computed over the last $|HTW|$ frames is stored in an output object data base noted DB_o.

In the case of a person (at frame t) referring to an unmatched mask in frame $t-1$, the unmatched person is initially labeled as 'get-in'. When a 'get-in' person is found, if there is no 'get-out' persons (DB_o is empty) we label this 'get-in' person as new. This 'get-in' person is then tracked along the video using the previously described protocol. Furthermore, if there is at least one 'get-out' person we should postpone the identification of this 'get-in' person by labeling it as 'unknown'. This 'unknown' person is then tracked on $|HTW|$ frames in order to have its description by a triplet (w, ρ, S). Using this description we calculate the value of kernel K_{change} (equation 8) between this unknown person and all get-out persons present in our database. Similarities between the unknown person and get-out persons are sorted in decreasing order so that the first get-out person of this list corresponds to the best candidate for a re-identification. Our criterion to map an unknown person to 'get-out', and thus to label it as get-back is based on a threshold on the first two maximal similarity values max_{ker} and max_2 of the list of similarities ($max_2 \leq max_{Ker}$). This criterion called, SC is defined as $max_{ker} > th_1$ and $\frac{max_2}{max_{Ker}} < th_2$, where th_1 and th_2 are experimentally fixed thresholds. Notice that, SC is reduced to a fixed threshold on max_{ker} when the set of get-out persons is reduced to one or two elements. An unknown person whose SC criterion is false is labeled as a new person. Both new and get-back persons are tracked between frames until they get out from the video and reach the get-out state.

Classically, any tracking algorithm has to cope with many phenomena such as occlusions. In this paper we limit the study to overlapping bounding boxes. When an overlap greater than an experimentally fixed threshold occurs between two bounding boxes, an occlusion is found. We assume two kinds of occlusions: partial occlusion where the occluded object remains visible and severe occlusion where the occluded object is completely hidden. If two or more objects (detected

at time t) merge together (at time $t+1$) to form one new object, this object is deemed to be a group rather than an occlusion. Group cases are not considered in this work.

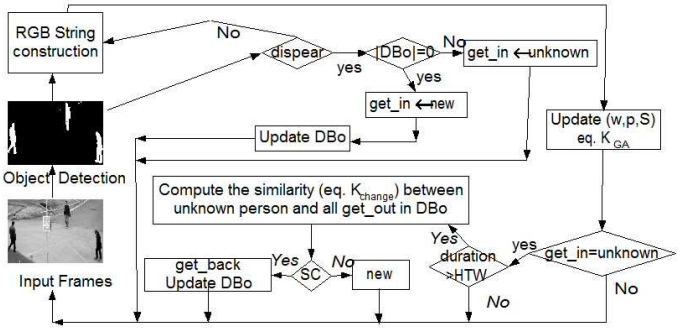

Fig. 3. The block diagram of the proposed method

5 Experiments

The proposed algorithm has been tested on v01 and v05 video sequences of the PETS'09 S2L1 dataset[1]. Each sequence contains multiple persons and occlusions cases.

In our first experiment we have evaluated how different values of the length of HTW may affect the re-identification accuracy. Figure 4 shows the effects for HTW changes on the true positive measurement for each view. The obtained results show that, v01 performs at peak efficiency for HTW=30. Video v05 attains its optimum at HTW=20. These curves also show that the length of HTW is not a crucial parameter of our method.

In a second experiment we show the improvement of the proposed kernel with respect to an histogram based approach. Similarly to [15] where histograms are defined from the already extracted blob (segmented parts), we propose the following histograms construction scheme: color histograms are computed on HTW frames for both the query object and each get-out persons contained in DB_o. Then, we try to map the query object with one of the get-out objects already stored in DB_o using EMD distance [14] between histograms. If a map is found, the query object gets the label of the mapped get-out object, and we update DB_o. Otherwise, we create a new label for the query object. The map criterion used here is similar to the above SC criterion (Section 4), nevertheless the best candidate corresponds to the minimum since we use distances. Table 1 reports the comparison results between histogram-based and kernel-based approaches.

[1] Available at http://www.cvg.rdg.ac.uk/PETS2009/a.html

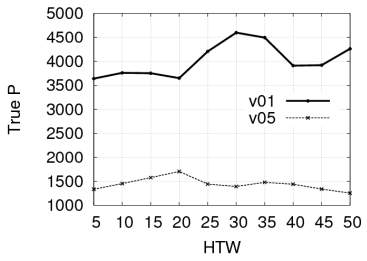

Fig. 4. HTW effects

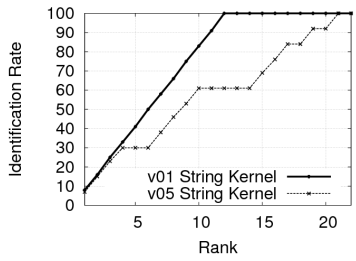

Fig. 5. CMC curves

As it can be seen from Table 1 the results of the proposed kernel give slightly higher values than the histogram approach regarding v05, while the results are clearly better for v01. We attribute this to the high detection accuracy in v01. Furthermore, v05 contains a lot of severe occlusions which are not specifically addressed in the proposed method. Indeed on such severe occlusions a large part of a person is usually hidden by an other person.

To validate our method of re-identification we used the Cumulative Matching Characteristic (CMC) curves. The CMC curve represents the percentage of times the correct identity match is found in the first n matches. Figure 5 shows the CMC curves for the two views. We can see that the performance of v01 is much better than that of v05. This last result being due to the large number of occlusions occurring in v05.

Table 1. Kernel vs. histogram

	view01		view05	
	Hist	Ker	Hist	Ker
True P	2514	4885	1360	1706
False P	2371	0	1365	1019
Missed	124	124	84	84

Table 2. Evaluation results

View	work of [1]	current work		
	MODA	MODA	MOTA	SFDA
v01	0.67	0.97	0.97	0.91
v05	0.72	0.60	0.60	0.81

We also used the exhaustive comparison of 13 methods defined in [7] in order to compare our results to the state of the art. The study [7] did a quantitative evaluation of the results submitted by contributing authors of the two PETS workshops in 2009 on PETS'09 S2.L1 dataset. We noticed that the submitted results of [1] outmatch all other methods using the MODA, MOTA, MODP, MOTP SODA and SFDA metrics described in [12]. Therefore, we only compare our results to this last method. Table 2 depicts the following: the left column shows the best results [1] obtained by methods described in [7] on each video. The second columns of Table 2 shows that our method obtains a lower MODA index than [1] on v05 but clearly outperform this last method on v01. These results may again be explained by the high number of occlusions in v05 which are overcome by [1] using multiple views of each person while the present method

is restricted to a single view. These results indicate thus the relevance of the proposed re-identification method when objects are not severely occluded.

6 Conclusion

In this work, we addressed the people re-identification problem by proposing a new approach based on RGB string kernels. Our re-identification system is based on a simple matching criterion to follow a person along a video. Person's description and kernel between these descriptions is used to remove ambiguities when a person reappears in the video. A benchmark public dataset was used to validate our method. Our experiments results show that the proposed approach outperformed state-of-the art methods when few severe occlusions occur.

Our future research will focus on the investigation of occlusion scene and group problems still using a single camera. To handle these phenomena, we should for each object severely occluded or entering into a group, suspend the update of its curves during the frames where it is hidden. Indeed in such cases no reliable feature may be extracted to characterize hidden persons.

References

1. Berclaz, J., Shahrokni, A., Fleuret, F., Ferryman, J., Fua, P.: Evaluation of probabilistic occupancy map people detection for surveillance systems. In: Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), pp. 55–62 (2009)
2. Brun, L., Conte, D., Foggia, P., Vento, M.: People re-identification by graph kernels methods. In: X. Jiang, M. Ferrer, A. Torsello (eds.) Graph-Based Representations in Pattern Recognition, *Lecture Notes in Computer Science*, vol. 6658, pp. 285–294. Springer Berlin Heidelberg (2011)
3. Cong, D.N.T., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People re-identification by spectral classification of silhouettes. Signal Processing **90**(8), 2362–2374 (2010)
4. Cuturi, M.: Fast global alignment kernels. In: L. Getoor, T. Scheffer (eds.) ICML, pp. 929–936. Omnipress (2011)
5. Desobry, F., Davy, M., Doncarli, C.: An online kernel change detection algorithm. IEEE Transactions on Signal Processing **53**(8-2), 2961–2974 (2005)
6. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. Journal of Ambient Intelligence and Humanized Computing **2**, 127–151 (2011)
7. Ellis, A., Shahrokni, A., Ferryman, J.M.: Pets2009 and winter-pets 2009 results: a combined evaluation. In: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. IEEE (2009). Conference held 7-9 Dec 2009 in Snowbird, USA.
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010). IEEE Computer Society, San Francisco, CA, USA (2010)

9. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: ICDSC, pp. 1–6. IEEE (2008)
10. Ijiri, Y., Lao, S., Han, T.X., Murase, H.: Human re-identification through distance metric learning based on jensen-shannon kernel. In: VISAPP (1), pp. 603–612 (2012)
11. Iodice, S., Petrosino, A.: Person re-identification based on enriched symmetry salient features and graph matching. In: J.A. Carrasco-Ochoa, J.F.M. Trinidad, J.S. Rodrguez, G.S. di Baja (eds.) MCPR, *Lecture Notes in Computer Science*, vol. 7914, pp. 155–164. Springer (2013)
12. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. Pattern Analysis and Machine Intelligence **31**(2), 319–336 (2009)
13. Mahboubi, A., Brun, L., Conte, D., Foggia, P., Vento, M.: Tracking system with re-identification using a graph kernels approach. In: R. Wilson, E. Hancock, A. Bors, W. Smith (eds.) Computer Analysis Images and Patterns 2013, vol. lncs 8047, pp. 401–408. York (2013)
14. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth moverś distance as a metric for image retrieval. International Journal of Computer Vision **40**(2), 99–122 (2000)
15. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: 16th Int. Conf. on Image Analysis and Processing (ICIAP 2011). Ravenna, Italy (2011)
16. Schwartz, W.R., Davis, L.S.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Brazilian Symposium on Computer Graphics and Image Processing, pp. 322–329 (2009)
17. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: ICCV, pp. 1–8. IEEE (2007)
18. Zhao, S., Precioso, F., Cord, M.: Spatio-temporal tube data representation and kernel design for svm-based video object retrieval system. Multimedia Tools Appl. **55**(1), 105–125 (2011)